MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 8: Difference-in-Differences 2

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2024

Course Outline

- **Week 1:** The potential outcomes framework
- **Week 2:** Randomized experiments
- **Week 3:** Estimation under selection on observables I
- **Week 4:** Estimation under selection on observables II
- **Week 5:** Estimation under selection on observables III
- *Week 6: Reading week*
- **Week 7:** Difference-in-differences I
- **Week 8:** Difference-in-differences II
- **Week 9:** Instrumental variables I
- **Week 10:** Instrumental variables II
- **Week 11:** Regression discontinuity

Difference-in-Differences This Far

So far we have considered the canonical 2-period difference-in-differences (DiD) design, with a brief foray into a special 3-period case.

Identification and estimation was reasonably straightforward:

- Key identification assumption is parallel trends, plus no anticipation
- Use either a plug-in or regression-based estimator

However, we often encounter DiD settings that are more complex:

- More than 2 time periods
- Treatment is assigned variably over time
- Treatment effects are heterogeneous (over time)
- Treatment is non-absorptive

Today we will consider identification and estimation in such settings.

Today

1. A(nother) Motivating Example
2. Fixed Effects Estimation
3. Variable Treatment Timing
4. Multi-Period Designs with Heterogeneous Treatment Effects
5. Synthetic Control Method Primer

Minimum Wage and Employment

Recall Card & Krueger (1994) used DiD to study the minimum wage (MW) policy in New Jersey, finding a positive effect of minimum wage on employment. Card & Krueger (2000) revisited this design and setting with better data, and found no effect either way.

Lots of debate – many papers reconsidered this question using a more general approach: Leveraging cross- and within-state variation throughout the USA. They largely find negative effects on employment.

Dube, Lester, and Reich (2010) revisit this debate:

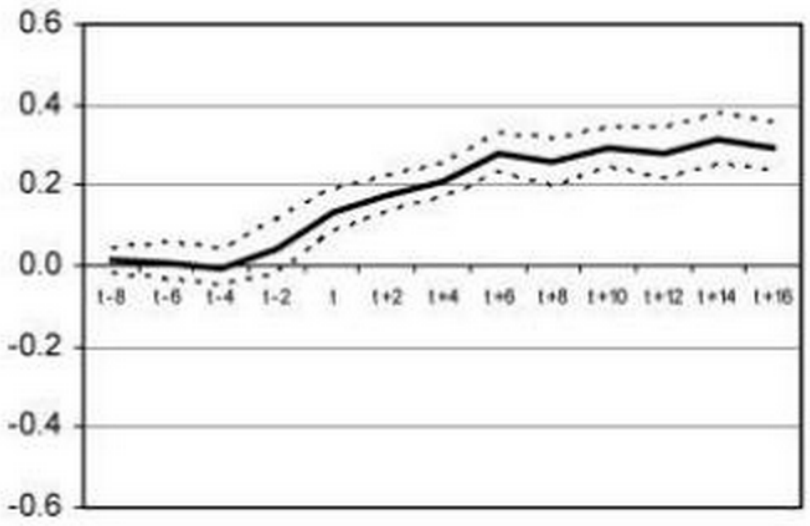
- Find all cross-state-border changes in MW policies (1990 - 2006)
- Collect earnings and employment data for every county in the USA in this time period.
- Generalize the DiD case study approach.

Variation in Space

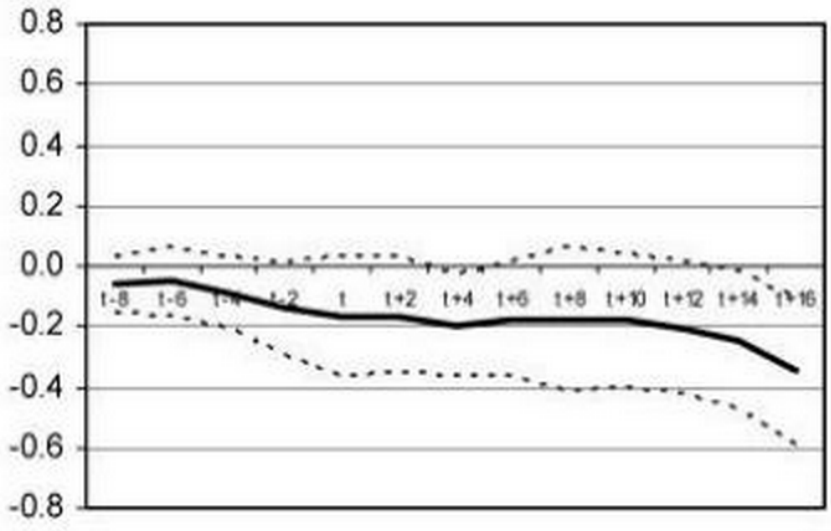
Estimated Dynamic Effects – Entire Sample

Ln Earnings

1. All County Sample, Common Period Effects

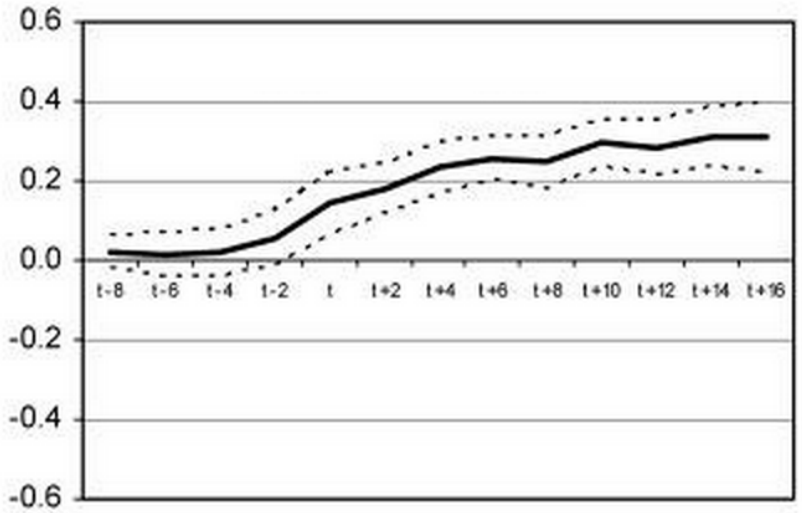


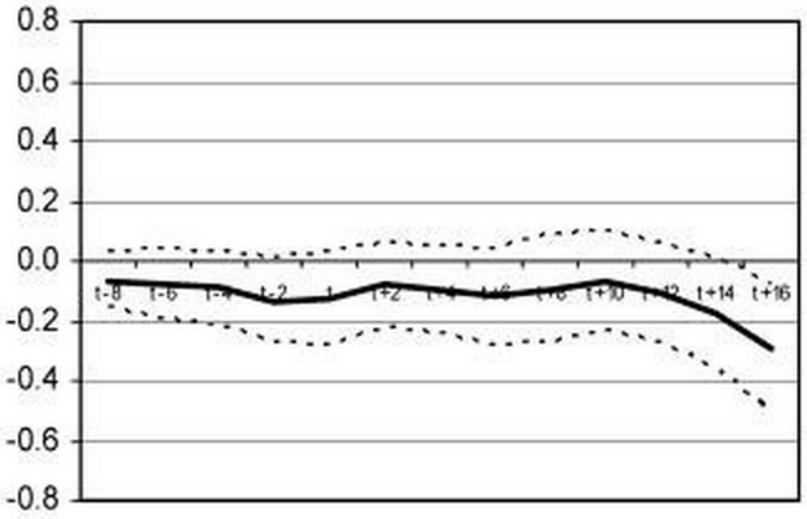
Ln Employment



Estimated Dynamic Effects – Border Sample Only

5. Contiguous Border County-Pair Sample, Common Period Effects





Fixed Effects Estimation and Difference in Differences

Recall the additive linear model for panel data with 2 periods:

$$Y_{it}(z) = \alpha_i + \gamma t + \tau z + \varepsilon_{it}$$

where

- $i \in \{0, ..., N\}$: Unit indicator
- $t \in \{0, 1\}$: Time indicator
- $Y_{it}(z)$: Potential outcome under treatment status $Z \in \{0, 1\}$
- α_i: time-invariant unobserved effect
- ε_{it}: idiosyncratic error term
- τ: (Homogeneous, constant) treatment effect of interest

In a 2-period design, we saw that the first-difference regression:

- Unbiasedly estimates τ when parallel trends and no anticipation assumptions satisfied
- τ will coincide with τ_{ATE} and τ_{ATT} if model is correct

Fixed effects estimation generalises this to $t > 2$

Panel Data Notation and Setup

y_{it}: Observed outcome for unit i in period t

$\boldsymbol{x}_{it} \equiv [z_{it}, x_{it1}, ..., x_{it(K-1)}]^{\top}$: Explanatory variables (including both treatment and observed covariates) for unit i in period i

We observe a sample of $i = 1, 2, ..., N$ cross-sectional units for $t = 1, 2, ...T$ time periods (a balanced panel)

Panel Data Notation and Setup

Collect variables for unit i:

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T \times 1} \qquad \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1}^\top \\ \vdots \\ \boldsymbol{x}_{it}^\top \\ \vdots \\ \boldsymbol{x}_{iT}^\top \end{pmatrix} = \begin{pmatrix} z_{i1} & x_{i11} & \cdots & x_{i1(K-1)} \\ \vdots & \vdots & & \vdots \\ z_{it} & x_{it1} & \cdots & x_{it(K-1)} \\ \vdots & \vdots & & \vdots \\ z_{iT} & x_{iT1} & \cdots & x_{iT(K-1)} \end{pmatrix}_{T \times K}$$

And stack them for all units (a "long panel"):

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_i \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix}_{NT \times 1} \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_i \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix}_{NT \times K}$$

Pooled OLS Model

When we ignore the panel structure and regress y_{it} on $\boldsymbol{x}_{it}$ we get

$$y_{it} = \boldsymbol{x}_{it}^{\top}\beta + v_{it}, \qquad t = 1, 2, ..., T$$

with composite error $v_{it} \equiv \alpha_i + \varepsilon_{it}$

$\hat{\beta_{OLS}}$ will be unbiased and consistent when:

$$E[v_{it}|\boldsymbol{x}_{it}] = 0 \text{ for } t = 1, 2, ..., T$$

i.e. $\boldsymbol{x}_{it}$ is strictly exogenous

- Read: the composite error v_{it} in each time period is uncorrelated with the past, current, and future regressors
- This is equivalent to strict conditional ignorability of potential outcomes.

Fixed Effects Model

Our unobserved effects model is:

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

We can estimate both $\boldsymbol{\beta}$ and α_i via OLS:

$$(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_N) = \underset{\boldsymbol{\beta}, \alpha_1, \ldots, \alpha_N}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} - \alpha_i)^2$$

$\hat{\boldsymbol{\beta}}$ is called the least squares dummy variables (LSDV) estimator

This is a generalization of the pre-post design we discussed last week

Fixed Effects Estimators

1. $\hat{\boldsymbol{\beta}}$ can be obtained via first differences estimation:
   a. Create differenced variables: $\Delta \boldsymbol{x}_{it} = \boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1}$ and $\Delta y_{it} = y_{it} - y_{i,t-1}$
   b. Regress Δy_{it} on $\Delta \boldsymbol{x}_{it}$
      Note: By taking the first difference we "purge" the fixed effects α_i
2. Can also be obtained via within estimation:
   a. Create demeaned variables: $\ddot{\boldsymbol{x}}_{it} \equiv \boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i$ and $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$
   b. Regress $\ddot{y}_{it}$ on $\ddot{\boldsymbol{x}}_{it}$
      Note: By within-demeaning we purge the fixed effects α_i
3. Or with LSDV estimation:
   a. Regress y_{it} on $\boldsymbol{x}_{it}$ and unit dummies
      Note: Here we directly estimate α_i, no purging

All 3 procedures are consistent with T fixed and $N \to \infty$ under the same assumptions.

Procedure 1 can be more efficient under serial correlation, while 2 and 3 are exactly equivalent in terms of point estimation.

When N is very large, 3 is computationally expensive and 2 preferred.

Fixed Effects Estimators: Assumptions and Uncertainty

Assumptions:

1. Strict exogeneity conditional on the unobserved effect
   - $\mathbb{E}[\varepsilon_{it}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}, \alpha_i] = 0, \ t = 1, 2, ..., T$
   - allows $\boldsymbol{x}_{it}$ to be arbitrarily related to α_i
   - SUTVA is implicitly assumed both across units and time periods
2. No carryover effects
   - Treatment status for any Z_{it} does not directly affect outcome $Y_{i,t'>t}$
3. Rank condition
   - Regressors vary over time for at least some i and are not perfectly collinear

Under these assumptions, $\hat{\beta_{FE}}$ is unbiased and consistent as $N \to \infty$
(But note that $\hat{\alpha}_i$ via LSDV is *inconsistent* for fixed T and $N \to \infty$)

Uncertainty Estimation:

- Usually SEs should be clustered by unit to account for serial correlation and clustered treatment assignment
- If N is small use block bootstrap

Adding Time Effects

Consider again our unobserved effects model:

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

Typical violation of strict exogeneity assumption: Common shocks that affect all units' y_{it} in the same way and are correlated with $\boldsymbol{x}_{it}$.

- Trends in farming technology or climate affect productivity
- Trends in immigration inflows affect naturalization rates
- Economic recession affects employment

We can allow for common shocks by including time effects:

- linear time trends
- non-linear time trends
- unit-specific linear time trends
- time fixed effects

Two-Way Fixed Effects Regression

Focus on time fixed effects:

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \alpha_i + \delta_t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

where

- α_i represents the unit effect
- δ_t represents common shocks in each time period

This is the two-way fixed effects (TWFE) model.

If our model is correct and $\boldsymbol{x}_{it}$ includes binary z_{it} (our DiD treatment indicator, taking 1 for the treated group in the post-period), then the TWFE is generalized difference-in-differences.

Use typical FE estimators (FD, within, LSDV) with both unit and time effects; in `R`:

- `lm` (slow!)
- `plm`
- `fixest`

Dynamic Two-Way Fixed Effects

We can specify a TWFE that allows for dynamic (time-varying) treatment effects:

$$y_{it} = \alpha_i + \delta_t + \sum_{r \neq 0} \mathbf{1}[R_{it} = r]\tau_r + \varepsilon_{it}$$

where

- α_i represents the unit effect
- δ_t represents common shocks in each time period
- R_{it} is the period relative to treatment for unit i
- τ_r is a relative-period treatment effect

This estimator, sometimes called the event study estimator, allows for heterogeneous treatment effects, but of a specific form: they cannot vary across treatment cohorts.

Variable Treatment Timing

Multi-period treatment regimes usually vary over two dimensions:

- Uniform vs. Staggered: Does treatment occur simultaneously, or over time?
- Absorbing vs. Non-absorbing: Once treatment occurs, can it switch off?

With anything other than uniform and absorbing treatment timing, TWFE for DiD may not behave well. For synthesis, see:

- Baker, Larcker, and Wang (2022)
- Roth, Sant'Anna, Bilinski, and Poe (2023)

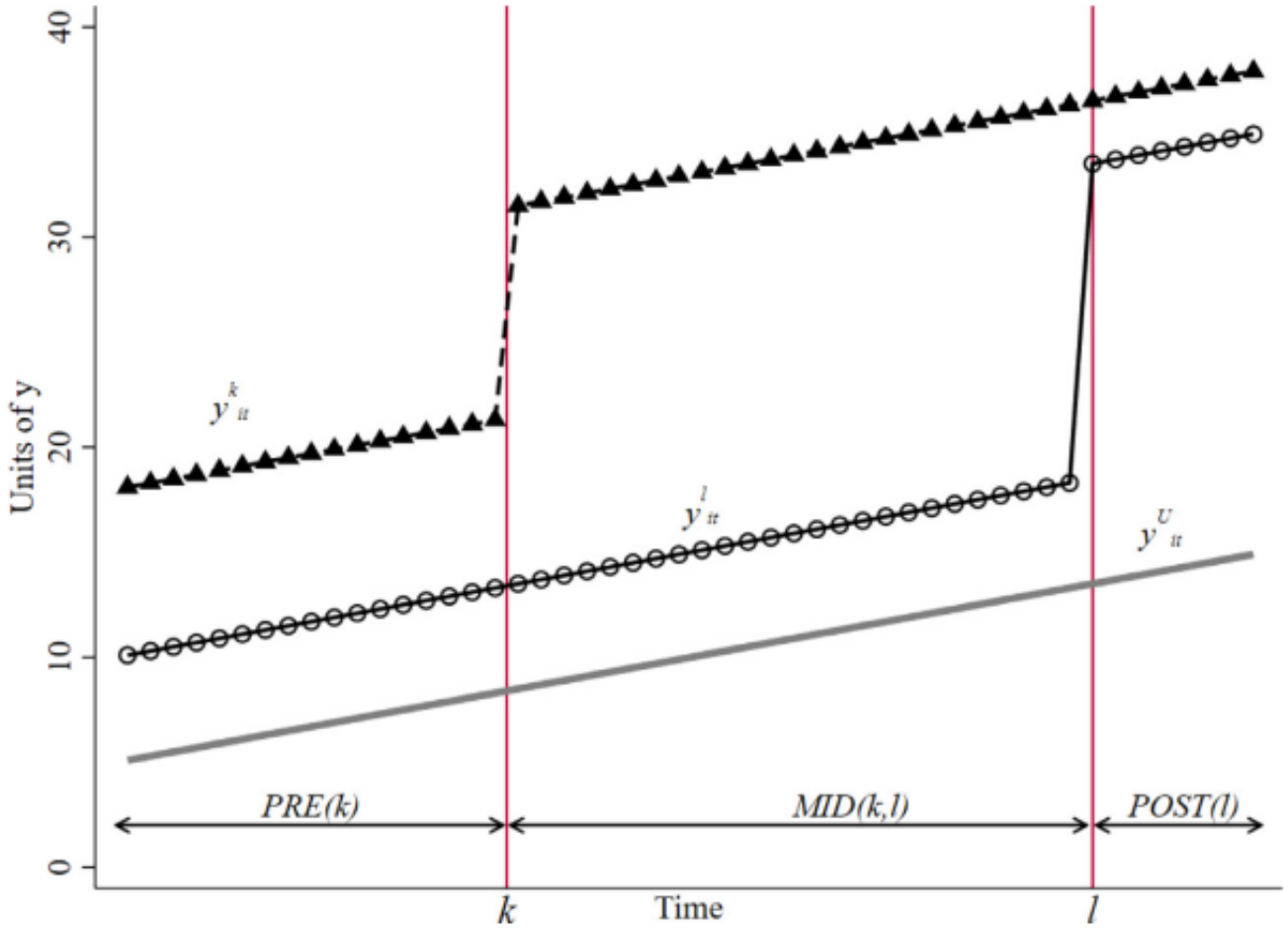
Short of further assumptions, the estimand targeted by TWFE is not easily interpretable $\rightsquigarrow$ it is a weighted average of many different treatment effects.

These weights can be negative (!), are generally non-intuitive, and can potentially severely mislead (e.g. sign-flips).

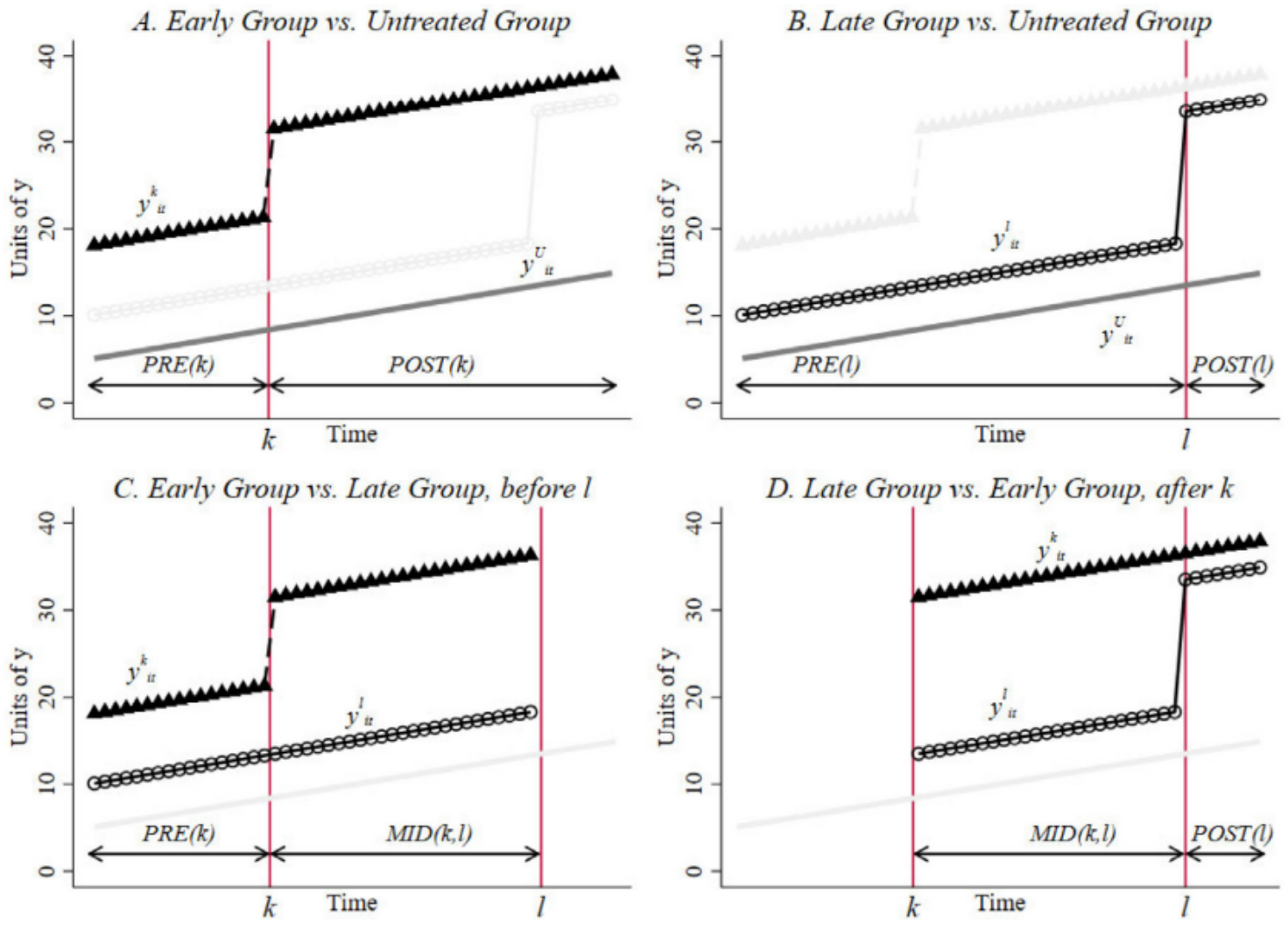
τ_{TWFE} Decomposition

To see this, we can decompose τ_{TWFE}. We focus on Goodman-Bacon (2021).

Define three groups: never treated (U), early treated (k), and late treated (l)



τ_{TWFE} Decomposition



τ_{TWFE} is the weighted average of these four 2x2 treatment effects.

τ_{TWFE} Decomposition

More generally:

$$\tau_{TWFE} = \sum_{g \neq g',t,t'} v_{g,g',t,t'} \tau_{g,g',t,t'}$$

where

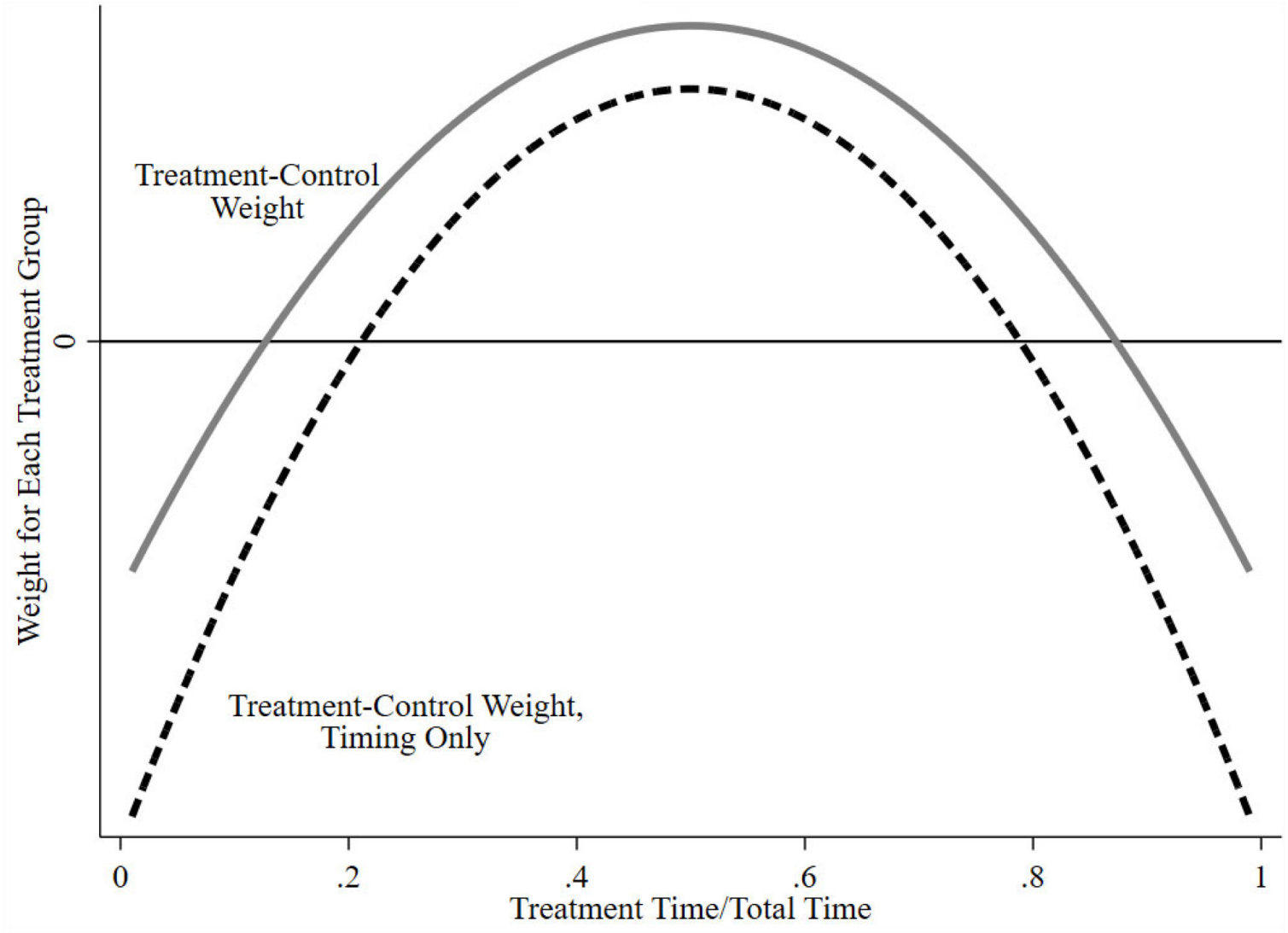
- $\tau_{g,g',t,t'}$ is the comparison of group g to group g', from time t to time t'
- $v_{g,g',t,t'}$ are weights that sum to 1

The weights are a function of the $N_{g,g',t,t'}$, $\frac{N_{1,g,g',t,t'}}{N_{0,g,g',t,t'}}$, and the relative timing of treatment.

Some of these comparisons may be 'forbidden': Already-treated units used as controls after they are treated. This is where negative weights arise.

τ_{TWFE} Decomposition

Weighting is heavily dependent on timing:



Key Assumptions for TWFE as Generalised DID

This decomposition reveals the assumptions under which traditional TWFE might be trusted with multi-period panel data and a DiD design.

A causal DiD interpretation with TWFE requires either:

1. Parallel trends, no anticipation, and homogenous τ

   Or:

2. Parallel trends, no anticipation, and uniform timing (constant τ over time)

Quickly explore how much of a problem this may be using `bacondecomp` in `R`.

Two Classes of Modern Estimators

Multi-period DiD with non-uniform (staggered or non-staggered) treatment timing should be approached with caution.

Two general types of modern estimators that can help:

1. Flexible matching and re-weighting estimators:
   - Make the 'right' comparisons only, weight appropriately, and recover τ_{ATT}.
   - Many estimators exist: Strezhnev (2018), de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021), Imai and Kim (2021), Dube et al. (2023), de Chaisemartin and D'Haultfœuille (2024)
   - We will focus on Callaway and Sant'Anna (2021)
2. Counterfactual estimators:
   - Estimate only Y_0, thus avoiding forbidden comparisons, and recover τ_{ATT}.
   - Many estimators exist: Gobillon and Magnac (2016), Xu (2017), Borusyak et al. (2021), Gardner (2021), Wooldridge (2021)
   - We will focus on Liu, Wang, and Xu (2022)

Callaway and Sant'Anna (2021): Setup

Callaway and Sant'Anna (2021) study a DiD setting with multiple time periods, staggered treatment timing, there may be heterogeneous τ, and parallel trends may hold only conditional on X.

They begin by defining a new estimand, the group-time ATT:

$$\tau_{g,t}^{ATT} = \mathbb{E}[Y_t(1) - Y_t(0)|G_g = 1]$$

where

- There are $T = t \in \{1, ..., T\}$ time periods
- $G_g \in \{0, 1\}$ indicates whether a unit is first treated in period g
- $Y_t(1)$ and $Y_t(0)$ are potential outcomes under treatment and control, for t

Intuitively, this has already brought us a long way. We can now reason in abstraction about every 2x2 comparison in our data.

Callaway and Sant'Anna (2021): Identification

This group-time ATT can be identified as:

$$\tau_{g,t}^{ATT} = \mathbb{E}\Bigg[\underbrace{\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)}_{\text{Weights}} \underbrace{(Y_t - Y_{g-1})}_{\text{Long difference in } Y}\Bigg]$$

where

- $C \in \{0, 1\}$ which takes 1 if never treated (no forbidden comparisons!)
- $p_g(X) = P(G_g = 1 | X, G_g + C = 1)$ is a propensity score
- Up-weight control units similar in $p_g(X)$ to the group-specific treated units

If parallel trends holds without conditioning on X, this simplifies to:

$$\tau_{g,t}^{ATT} = \mathbb{E}[Y_t - Y_{g-1} | G_g = 1] - \mathbb{E}[Y_t - Y_{g-1} | C = 1]$$

Callaway and Sant'Anna (2021): Estimation

Estimation proceeds as follows:

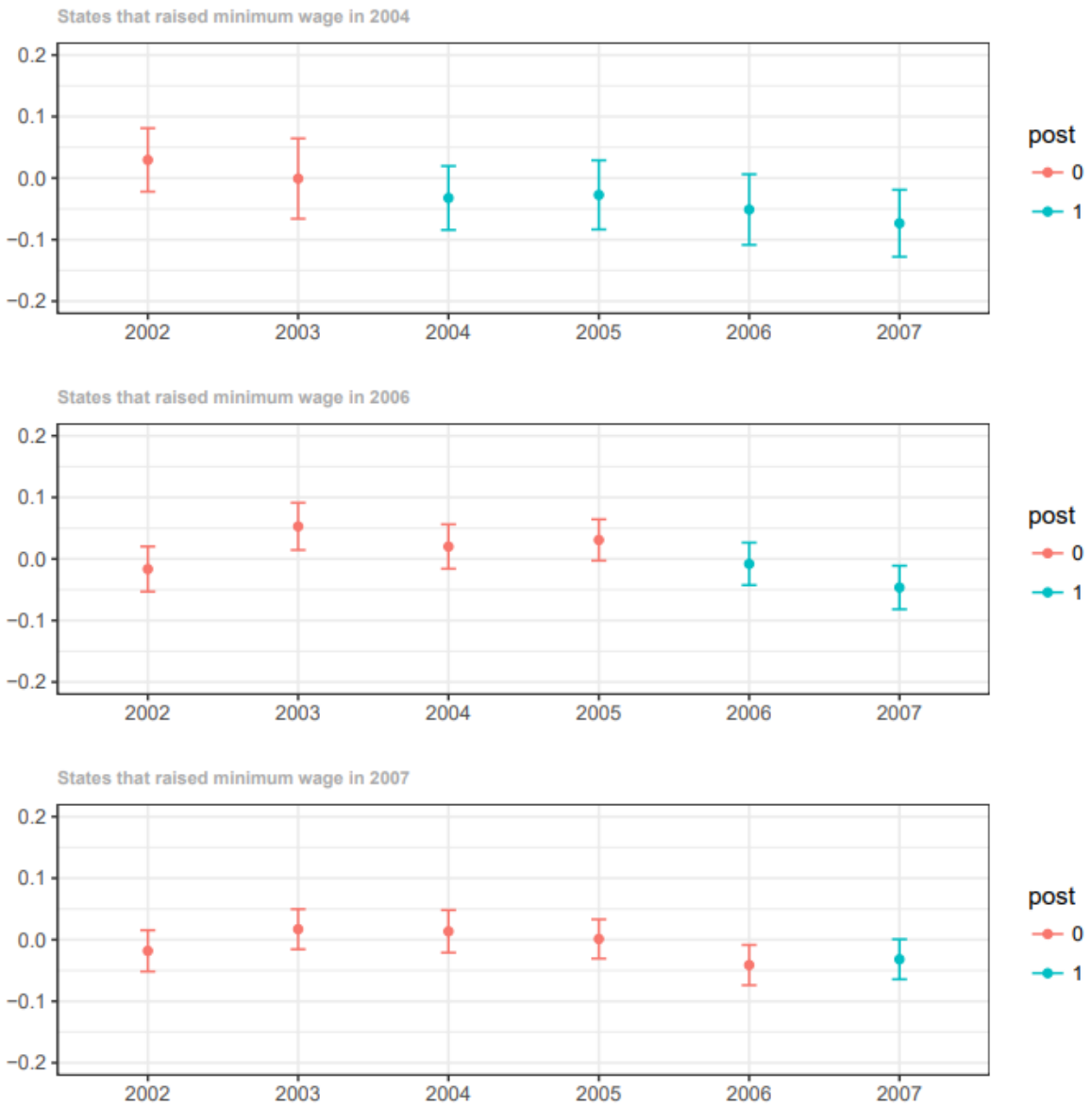
1. Estimate $\hat{p_g}$ for each group g
2. Estimate $\tau_{g,t}^{\hat{ATT}}$ by plugging in fitted values and observed Y into the (estimator-version) of the expression on the previous slide
3. Combine the estimated values of $\tau_{g,t}^{\hat{ATT}}$ to retrieve quantities of interest

Some quantities of interest:

- Simple average of $\tau_{g,t}^{\hat{ATT}}$ across t and g
- Weighted average of $\tau_{g,t}^{\hat{ATT}}$ weighting by group sizes
- Any other principled summary measure!

All this can be done in `R` with package `did`

Callaway and Sant'Anna (2021) on the Minimum Wage



Liu, Wang, and Xu (2022): Setup

Liu, Wang, and Xu (2022) consider DiD settings with multiple time periods, staggered treatment timing that may or may not be absorbing, and there may be heterogeneous τ.

Define the estimand of interest as:

$$\tau_{ATT} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid Z_{it} = 1, C_i = 1]$$

where

- Z_{it} is our normal DiD treatment indicator
- C_i is an indicator for 'ever treated' units
- $Y_{it}(1)$ and $Y_{it}(0)$ are potential outcomes under treatment and control

Idea: Estimate only $Y_{it}(0)$ using pre-treatment data, taking $Y_{it}(1)$ as missing.

Estimate τ_{ATT} by taking differences between $Y(1)$ and $\hat{Y(0)}$.

Note: This is a philosophical departure from TWFE! Closely connected to synthetic control method and friends.

Liu, Wang, and Xu (2022): Estimation

Authors offer three estimators:

- FEct Estimator:

$$Y_{it}(0) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \alpha_i + \boldsymbol{t}_t + \varepsilon_{it}$$

- IFEct Estimator:

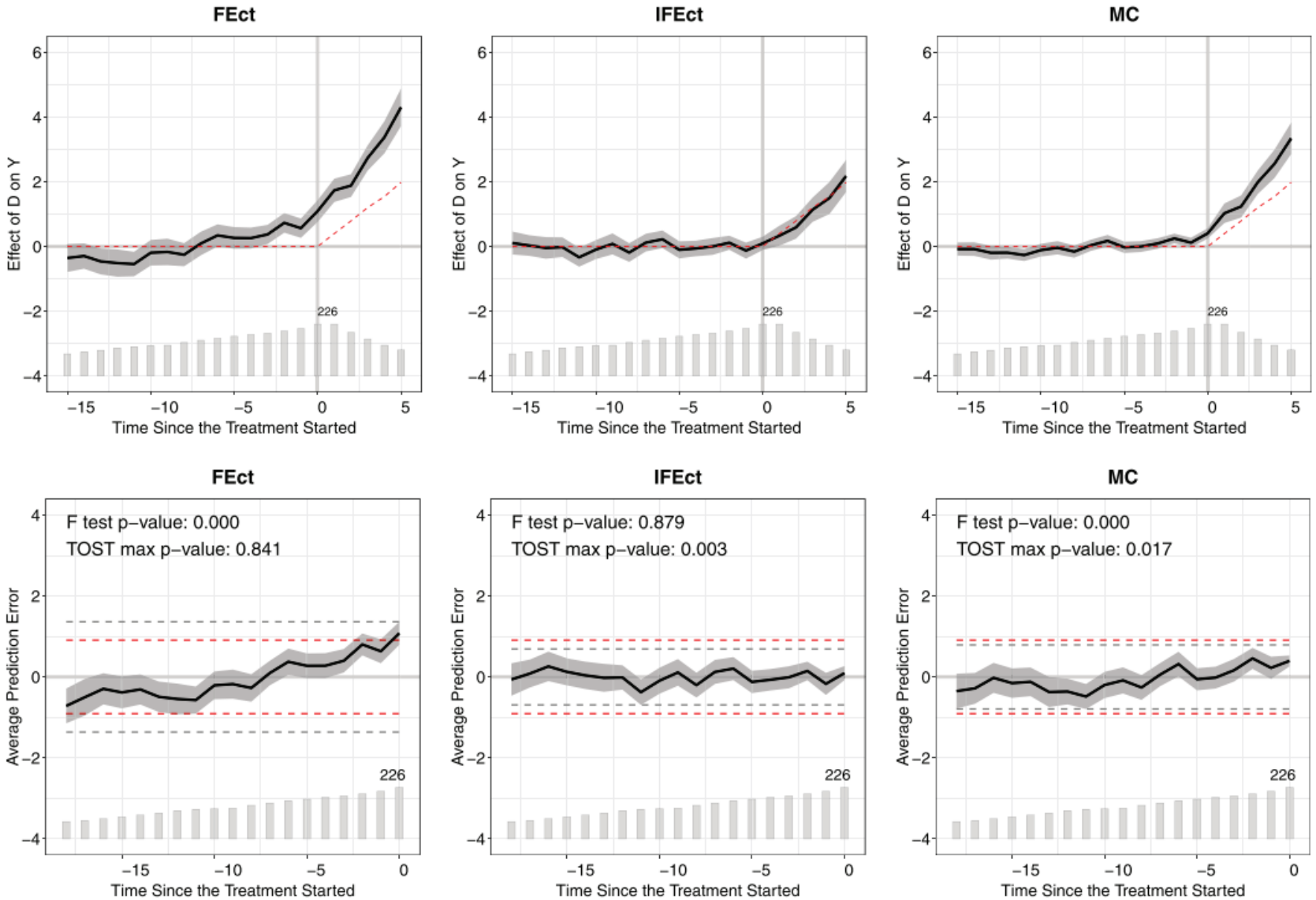
$$Y_{it}(0) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \alpha_i + \boldsymbol{t}_t + \lambda_i^{'}\boldsymbol{f}_t + \varepsilon_{it}$$

- MC Estimator:

$$\mathbf{Y}(\mathbf{0}) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \mathbf{L} + \boldsymbol{\varepsilon}$$

All this, plus diagnostics, can be done in `R` with package fect

Liu, Wang, and Xu (2022): Simulated Example



Synthetic Control Method: Basics

DiD requires parallel trends in the expected value of potential outcomes.
Generally cannot help where there are time- and unit-varying confounders.

Synthetic Control Methods (SCM) take a different approach:

1. Find $\mathbf{W}^*$ from the set of all $\mathbf{W}$, each of which are $N-1$ length vectors of unit weights, that minimizes $||\mathbf{X_1} - \mathbf{X_0}w||$ for $\mathbf{X_1}$ a matrix of pre-treatment outcomes and covariates for the treated unit, and $\mathbf{X_0}$ likewise for the control.

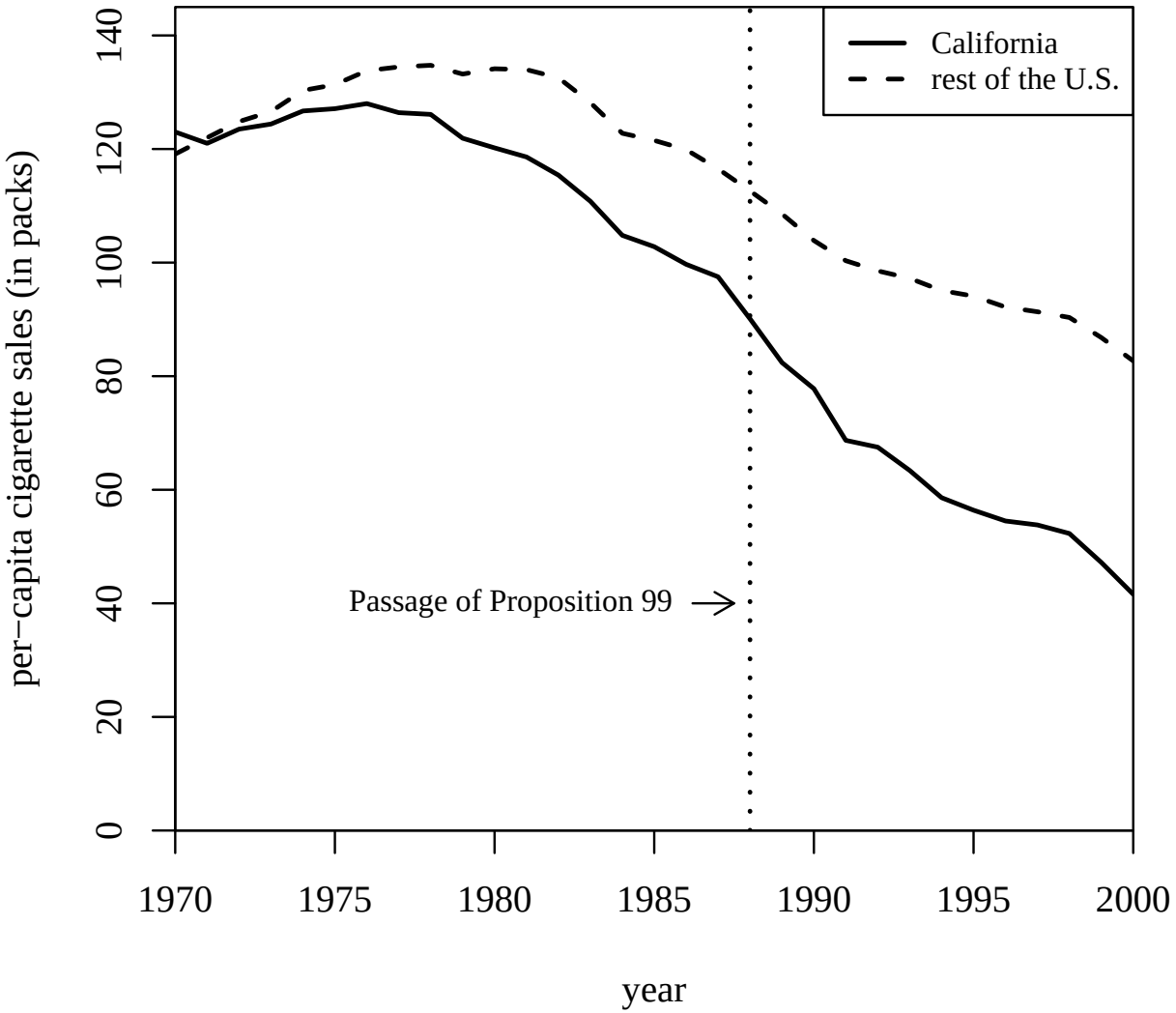
   $\rightsquigarrow$ That is, the weights $\mathbf{W}^*$ minimize the difference between t control units so that they match – in both levels and trends – the treated unit in the pre-treatment period. This weighted set of control units is the synethic control.

2. An approximately unbiased estimator of the unit-specific effect in the post-treatment period is:

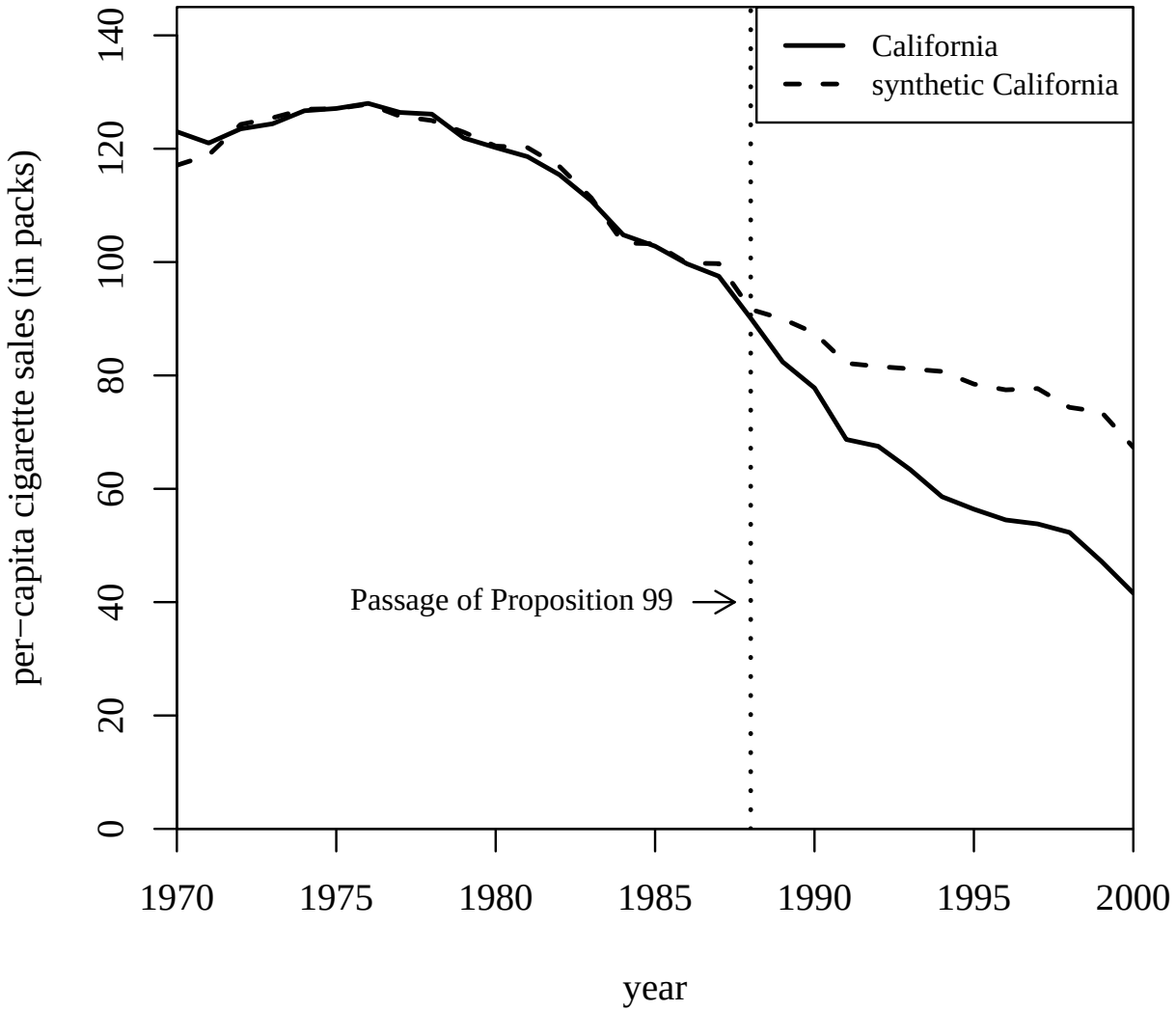
$$\widehat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \quad \text{for} \quad t \in \{T_0 + 1, \ldots, T\}$$

3. Inference uses placebos over time and across units (roughly, randomization inference).

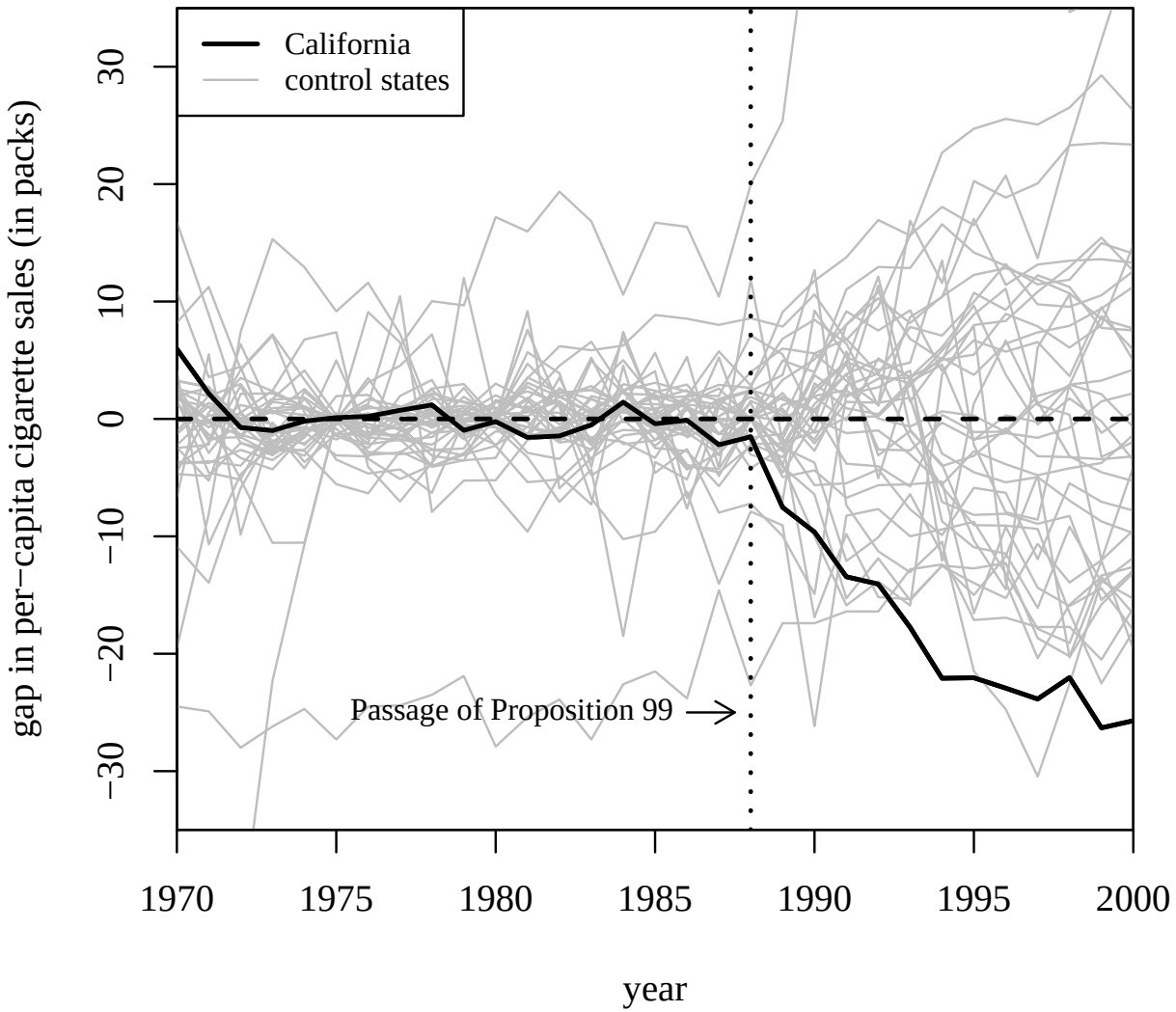
SCM Example: California Prop 99 (Abadie et al. 2010)



SCM Example: California Prop 99 (Abadie et al. 2010)



SCM Example: California Prop 99 (Abadie et al. 2010)



Into the Future

SCM is robust to time-varying confounding, so long as pre-trend fit is 'sufficiently close' $\rightsquigarrow$ intuition is that by matching on both pre-levels and pre-trends, any residual time-varying confounders are also matched. This is nice, but can be quite non-transparent.

For primer on use, see Abadie (2020).

Note that SCM was originally built to be used in single case studies.

New work has borrowed from both SCM and DiD to generalize this to multiple treated units – weighting control units *a la* SCM while estimating a DiD:

- Xu (2017): Generalized Synthetic Control Method (pre-cursor to fect suite)
- Ben-Michael et al (2018): Augmented Synthetic Control Method
- Arkhangelsky et al (2019): Synethic Diff-in-Diff

Summary

DiD (and friends) becomes more tricky with variable treatment timing:

1. TWFE estimation, long thought to be a simple generalisation of DiD, is an unintuitive re-weighting of various treatment effects
2. Cannot be ignored unless we are willing to assume constant treatment effects

Fortunately, new approaches exist:

- Flexibly estimate every comparison and re-weight/re-combine appropriately.
- Focus on only estimating the missing counterfactual Y_0.
- These approaches will also allow for more honest testing of pre-trends (*a la* Roth), but don't solve the pre-test selection problem!

Remember, however:

- The problems diagnosed here are theoretical, and don't always apply
- Often TWFE will give very similar estimate to more modern approaches
- TWFE is 'safe' with uniform timing of treatment